Ordinal logistic regression

Pamela Warner

What is it?

When a response variable has only two possible values (e.g. recurrence/not), binary logistic regression is commonly used to test or model the association between that response and a number of potential explanatory variables, with each association estimated in terms of an odds ratio (OR). Multinomial logistic regression is an extension of this approach to situations where the response variable is categorical and has *more than two possible values*. Ordinal logistic regression is a special type of multinomial regression, which can be advantageous when the response variable is ordinal. [See Box 1 for glossary of terms.]

When/why is it useful?

Ordinal response variables are common in medical research, but for the purposes of analysis it is often the case that such a variable will be recoded to just two levels, in order to be able to use standard binary logistic regression. For example, a pain score might be recoded to 'some pain' versus 'none', or 'severe pain' versus the rest. However, this strategy risks both data dredging (to 'select' the recode to be used) and loss of valuable information in the data.

Alternatively, standard multinomial logistic regression could be applied. This, however, requires fitting of a large number of parameters, so the number of degrees of freedom used in the model-fitting process can make excessive demands on the dataset. Furthermore, reporting/interpretation might have to be complex/lengthy. More problematically, multinomial logistic regression does not take into account the ordinal nature of the response variable, and so the analytic power to detect an association with an explanatory variable will be suboptimal when the association with response variable is effected evenly across the range of its possible values, namely when the entire distribution of responses is shifted along its ordinal scale (either up for a positive association or down).

Ordinal regression, however, is geared to response variables with ordinal effect, since it works with the cumulative distribution for the response variable, and the parameter it fits for each association represents the general trend across the ordinal values of the response variable.

Example of technique

A demonstration dataset has been created to illustrate these notes, where the ordinal response variable is effectiveness of contraception used, which for simplicity has three levels, from 'none used' to 'high' (whereas in the O'Rourke *et al.*[1] article in this issue the very similar outcome variable has four levels). Figure 1 shows this *artificial* dataset: 'effectiveness of contraception used', by two explanatory variables, namely gender and age group (those using no contraception are not shown but would make the columns total 100%). Modelling these data by standard multinomial regression, the ORs for the association with gender and age group would be as shown in the left panel of Figure 2. For the explanatory variables, the reference category for gender is female and for age is ≥30 years. For the response variable, the reference value is specified as 'no contraception', and association is estimated for each of its other values against this reference value. Therefore in this example, two sets of parameters are estimated, and in each 'sub-analysis' only a subset of the dataset is involved. It can be seen that all associations estimated for the response comparison 'moderate vs none' were found to be non-significant, whereas there were statistically significant associations for the more extreme comparison of 'high vs none'. Note that it is unknown to the analysis that 'moderate' is intermediate between none and high.

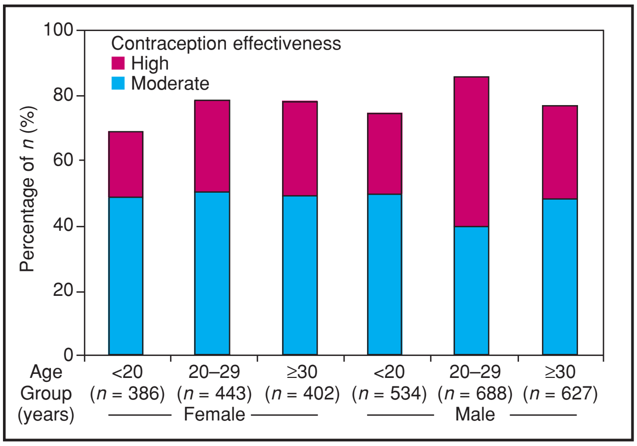

Figure 1 Effectiveness of contraception used, by gender and age group (artificial data)

In contrast, ordinal regression, using the entire dataset and taking into account the ordinal nature of contraception effectiveness, finds that there is a statistically significant trend for men to use more effective contraception than women *across the range of effectiveness levels* (high/moderate rather than none, *and* high rather than moderate/none). Similarly, for age group there are trends across the levels of the response variable, albeit in different directions for the two younger age groups (relative to the oldest). The ordinal logistic regression estimates can be seen to have narrower confidence intervals than those by multinomial logistic regression, confirming the greater power of this method for the same overall dataset. As would be expected, the ordinal OR estimates lie between the corresponding separate estimates by multinomial regression.

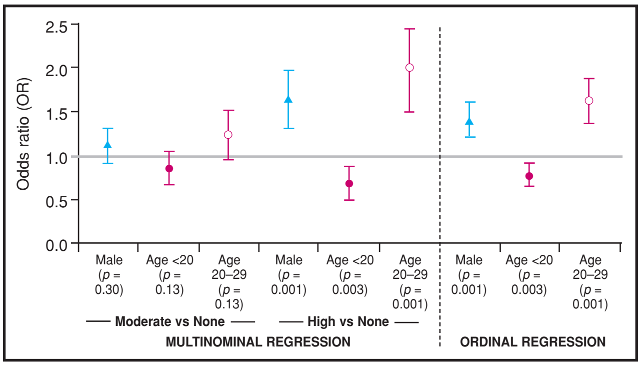

Figure 2 Estimates of associations with effectiveness of contraception used [odds ratio (OR), 95% confidence interval], using multinomial and ordinal logistic regression (artifical data)

J Fam Plann Reprod Health Care 2008; **34**(3): 169–170

Public Health Sciences, University of Edinburgh Medical School, Edinburgh, UK
Pamela Warner, BSc, PhD, *Senior Lecturer in Medical Statistics and Associate Editor, Journal of Family Planning and Reproductive Health Care*

Correspondence to: Dr Pamela Warner, Public Health Sciences, University of Edinburgh Medical School, Teviot Place, Edinburgh EH8 9AG, UK. E-mail: p.warner@ed.ac.uk

Box 1: Glossary of statistical terms used in this article

Binary variable This is a categorical variable with only two possible values (e.g. case or non-case). Also termed **dichotomous** or **binomial.**

Categorical variable Such a variable has a limited number of distinct values, which might be non-quantitatively descriptive (e.g. hair colour).

Cumulative distribution For each value of an ordinal or continuous variable, the cumulative frequency presents the accumulated number of occurrences for that or any ‘lower’ value (rather than indicating simply the frequency of occurrence of that value alone, as in an ordinary frequency distribution). The cumulative frequency for the last (highest) value must therefore encompass the entire sample. Cumulative frequency is often reported not as counts but as a percentage of the total sample, in which case the last value must have a cumulative frequency of 100%.

Degrees of freedom This can be thought of as the modelling capacity (or independent elements of information) in the dataset.

Logistic regression This is a method for analysis of the occurrence or not of a particular response value, in relation to potential explanatory variables. What is modelled for each combination of explanatory variables is the logarithm of the odds of that response value (which is termed a logistic transformation). Each association in the model is summarised/estimated in terms of an odds ratio (OR).

Multinomial variable This is a categorical variable with more than two possible values. Also termed **polychotomous**.

Odds (of a specified response) This is the number of occurrences of that response value divided by the number without that value (e.g. cases divided by non-cases).

Odds ratio (OR) for a specified response For a binary explanatory variable, the OR is calculated as the odds of the specified response in those with the explanatory feature, divided by the odds of that response in those without the explanatory feature. If there is truly no association then the two odds should be approximately equal and the OR approximately 1, which is therefore the value for a ‘null association’.

Ordinal variable This is special semi-quantitative type of categorical variable where the values are conceptually ordered, such as degree of pain (e.g. none, mild, moderate, severe) or effectiveness of contraceptive method used (e.g. none, moderate, high).

Parameter This is a component in the model, which needs to be estimated from the data, and doing so uses up available degrees of freedom. [The number of parameters needed for a multinomial regression model is a multiple of the number needed for a binary logistic regression model.]

Power This term is used here, loosely, as the probability of detecting from the study data what is in fact the real situation.

Reference category The category within a categorical explanatory variable that is chosen as the comparator for calculation of ORs (i.e. denominator odds).

Reference value This is the value within a *response* variable that is used as a comparator response. [Note that this is specified only in multinomial regression because in binary logistic regression there are only two possible response values, and so the reference value can be assumed to be the only other possible response value.]

Response variable The outcome or dependent variable that is to be modelled/tested.

In this example only categorical explanatory variables have been used. However, where the effect of an explanatory variable is linear it can be entered as a linear variable, as is the case in binary logistic regression, with a resultant saving in the number of parameters to be fitted (as done by O’Rourke *et al*.[1] for age and pregnancy intention).

Interpretation of findings

In general, it is best to interpret the ORs as indicating the relative odds of a higher level response, for the value of the explanatory variable under consideration, relative to its reference category. For men the odds of using more effective contraception is greater by 40% than women (with the 95% confidence interval for increase in odds being 23% to 61%). As regards age group, the increase in odds of using more effective contraception is about 60% for those aged 20–29 years, relative to those aged 30 years and over. This can be confirmed by examining Figure 1, where it can be seen that this age group reported greatest use of contraception at all (some vs none) and relatively more use of effective contraception (high vs moderate/none). Conversely, the youngest age group had decreased odds of effective use of contraception, relative to the oldest age group, by about 20%.

What precautions are needed?

The benefits of a special statistical method will accrue only if it is suited to the research question of interest, and the data comply with the assumptions required for that method. Furthermore, the usual caveats apply, as for general application of statistical methods, and for logistic regression modelling in particular. If there is an ordinal association then it can be envisaged that there should be an overall shift in the entire distribution of responses for a specific value of an explanatory variable compared to its reference category. The ordinal response values can be thought of as representing a set of thresholds along the range of possible values at which comparisons can be made of the cumulative frequencies above and below the threshold. Therefore, the key assumption needed for ordinal regression is of *proportional odds*, which can be thought of as parallel cumulative frequency distributions for any two subgroups for which association is being estimated (e.g. males and females). When this applies for a particular explanatory feature then the ‘threshold’ ORs can be assumed to be constant along the range, so that only one OR (parameter) needs to be estimated, which is the approach taken by ordinal logistic regression.

Overview

The fact that ordinal logistic regression fits just one parameter per association, regardless of the number of levels of the response variable, reduces the number of parameters that have to be fitted, thus increasing power and simplifying reporting and interpretation. If the research interest is in effects (association with the explanatory variables) *across the range of possible response values,* then provided the data and underlying true association are suited to the method, ordinal logistic regression[2] is the analytic method of choice, because it provides a more succinct representation and more powerful testing of the associations at work.

Statements on funding and competing interests

Funding None identified.

Competing interests None identified.

References

1 O’Rourke K, Richman A, Roddy M, Custer M. Does pregnancy/paternity intention predict contraception use? A study among US soldiers who have completed initial entry training. *J Fam Plann Reprod Health Care* 2008: **34**: 165–168.

2 Norusis MJ. Ordinal regression (Chapter 4). *Statistical Package for the Social Science (SPSS) 13.0 Advanced Statistical Procedures Companion.* 2005. http://www.norusis. com/pdf/ASPC_v13.pdf [Accessed 8 May 2008].